Stereo Image-based Relocalization for Robust Visual Odometry

Yusuke E.[1*], Masafumi N. [1]

[1]*Shibaura Institute of Technology, 3-7-5, Toyosu, Koto-ku, Tokyo 135-8548, Japan*

*ah20034@shibaura-it.ac.jp

Abstract: *Mobile mapping systems (MMS) and unmanned aerial vehicles (UAVs) are widely used to quickly and safely collect 3D data for inspecting infrastructure, such as bridges, dams, roads, and railroads. One technical challenge is that self-position estimation by visual odometry is not easy when images are blurred due to camera movement and rotation. Therefore, we focused on visual odometry issues when mounting autonomous mobile robots such as indoor flying UAVs. However, technical issues with visual odometry include the difficulty estimating self-position when images are blurred during camera movement and rotation. Therefore, we proposed a methodology that identifies visual odometry errors and directs users to the restarted position. Our methodology also restarts visual odometry using image matching on a sequence of images.*

Keywords: *Visual odometry, image matching, motion blur, odometry reinitialization*

Introduction

Mobile mapping systems (MMS) and unmanned aerial vehicles (UAVs) have been widely used to safely and quickly acquire 3D data on bridges, dams, roads, railways, and other structures. In particular, there has been an increase in cases where inspection images and point clouds are acquired using simultaneous localization and mapping (SLAM) in recent years. SLAM can be broadly classified into two types: SLAM using a laser scanner (LiDAR-SLAM) and SLAM using a camera for self-localization (visual odometry) as well as visual SLAM. In our study, we focused on the technical challenges of integrating visual odometry functionality into autonomous mobile robots such as indoor UAVs. Visual SLAM and visual odometry are intended for the control of autonomous robots and UAVs in non-GNSS (global navigation satellite system) positioning environments. However, a one technical issue with visual odometry is that it is difficult to estimate our's position with visual odometry when images are blurred due to camera movement or rotation. Conventional research in seamless indoor-outdoor UAV navigation has focused on visual simultaneous localization and mapping (Visual SLAM) integrated with the Robot Operating System (ROS), as well as 2D modeling and orthoimage generation of complex structures using UAVs equipped with OpenREALM. Studies utilizing ROS-based frameworks have typically conducted comparative analyses of sensing modalities, including LiDAR, monocular RGB cameras, and stereo camera systems. In our previous work, we developed flight control algorithms designed specifically for UAV-based infrastructure inspection tasks. Additionally, we proposed a methodology to enhance the stability of visual odometry by incorporating multidirectional inertial measurements in conjunction with stereo imagery. Furthermore, we introduced a seamless positioning approach that enables seamless transitions between visual odometry and RTK-GNSS modes. This ensures continuous and reliable localization in both GNSS-available and GNSS-denied environments. Therefore, in our study, we propose a method for determining visual odometry errors using image matching on consecutive images. This methodology guides the user to the restart position, and restarts visual odometry.

First, we describe our methodology, which consists of estimation relative camera pose by visual odometry and image matching for the camera pose, correction process. Next, we describe an overview of the indoor walking measurement experiments. In these experiments, we summarize and discuss data from two experimental sites obtained by visual odometry using an IMU stereo camera at two experimental sites. Then, we investigate whether errors in camera pose estimation by visual odometry during sharp turns are compensated for by 3D map matching. We also conducted verification experiments using 3D movement.

Literature Review

In recent years, unmanned aerial vehicles (UAVs) have become increasingly smaller and lighter, while facing stricter constraints on computing resources and power consumption. As a result, there is a growing demand for technologies that enable self-localization and environmental mapping independently of Global Positioning System (GPS). Under these limitations, the development of SLAM systems that can operate capable of operating with high accuracy in real time has become a critical challenge.

The study employs bearing observations to estimate the position of the UAV position during flight while simultaneously building a feature map. To enable undelayed feature initialization, the inverse depth methodology is applied. To enhance the robustness of data association, a combination of the Mahalanobis distance and scale invariant feature transform (SIFT) descriptor matching is adopted. We evaluated the performance of the proposed system was evaluated through both simulations and real-world experiments. The results demonstrate that the system can suppress vehicle position estimation errors when constructing a 3D feature map in non-GNSS environments.

One remaining challenge lies in the computational complexity of the extended Kalman filter (EKF)-based algorithm. When the state vector is augmented with newly observed features, the computational cost increases quadratically with the number of features. This is an unavoidable limitation of EKF-based real-time SLAM for small UAVs. Future work should therefore explore developing a Rao-Blackwellized FastSLAM framework and stabilization techniques that achieve linear computational complexity.

In another study, Saito et al. (2024) identified two major issues in UAV operations for bridge inspection. First, no current UAV can transition seamlessly between indoor and outdoor environments, such as those around bridges, where GNSS and non-GNSS positioning conditions alternate. This limitation poses significant challenges to flight control under bridges, where satellite signals are unavailable. Since autonomous UAV flight typically relies on GNSS positioning, seamless indoor-outdoor flight unachievable in such environments. Second, for multi-temporal inspections, survey images must be precisely superimposed. However, it is difficult to reproduce the exact camera position and viewing angle from a previous survey is difficult when using UAVs.

To address these issues, the study presents a seamless indoor-outdoor positioning method that integrates RTK-GNSS positioning and visual odometry. This approach allows for mode switching and suppressing mitigates the accumulation errors inherent to visual odometry. In addition, the study introduces a methodology for aligning the position and attitude of the UAV-mounted camera with those of the previous survey by detecting and controlling image feature points. This approach enables the accurate overlay of multi-temporal images through feature point detection and correspondence processing. Prototypes that incorporate these functions have been developed, and performance verification is underway. Future work will focus on developing a system that can use the high-resolution images acquired with this methodology as data for building information modeling (BIM) and civil information modeling (CIM).

Finally, we will present a paper on SLAM that does not use feature points. The reason for introducing this paper is that I plan to conduct research in environments where feature points are difficult to obtain, such as on the Lunar surface is a direct SLAM methodology that estimates camera motion and a semi-dense depth map by directly using high-contrast pixels in images, unlike conventional feature point-based methods methodologies. One major advantage is that it does not require feature point extraction or matching. This makes, it applicable even in environments with few corners or textures. For depth estimation, it combines static stereo with a fixed baseline with temporal stereo based on camera motion.

This prevents scale drift while accommodating a wide range of scenes. It also incorporates corrections for lighting changes, demonstrating high robustness in real-world environments. Evaluations using KITTI and EuRoC datasets have achieved real-time, high-precision results, making establishing it as a viable alternative to feature-based methodologies.

Methodology

The proposed methodology consists of three processes: self-position estimation with visual odometry, error detection with visual odometry, and visual odometry restart, as shown in Figure 1. First, relative camera poses are continuously obtained with visual odometry. Next, assuming errors occur in visual odometry due to image blurring during sharp turns, the system determines whether a sharp turn is occurring by utilizing the discontinuity between feature point detection quantities and camera pose estimation quantities. Furthermore, with image matching based on feature points, the system guides the system to the estimated restart position and resumes visual odometry.

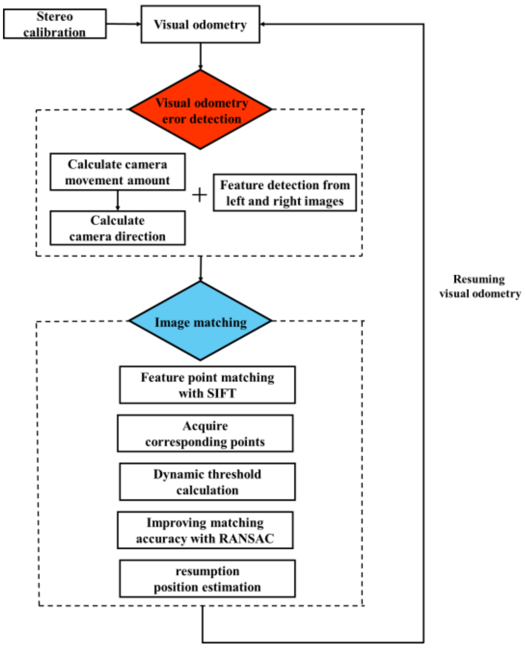
Stereo calibration
Visual odometry
Visual odometry eror detection
Calculate camera movement amount
Calculate camera direction
+
Feature detection from left and right images
Resuming visual odometry
Image matching
Feature point matching with SIFT
Acquire corresponding points
Dynamic threshold calculation
Improving matching accuracy with RANSAC
resumption position estimation

Figure 1: Proposed Methodology.

a. Stereo camera calibration:

In our study, we determine the intrinsic and extrinsic parameters of the stereo camera system by applying Zhang's widely used calibration methodology. This procedure, estimates camera parameters, such as the baseline distance between the two cameras, the focal lengths of each lens, and the lens distortion. Accurate calibration of these parameters is essential because they directly impact the precision of subsequent 3D reconstruction and localization tasks. To evaluate the reliability and accuracy of the estimated parameters, we compute the reprojection errors using images of a checkerboard pattern captured from multiple viewpoints. This evaluation provides quantitative evidence of the calibration quality and confirms that the stereo camera model is suitable for our subsequent experiments.

b. Visual odometry:

Visual odometry is the process of estimating a camera's pose by analyzing sequential images. For this technique to operate effectively, the surrounding environment must have enough variation in brightness and adequate surface textures, to reliably extract feature points throughout the motion estimation process. It is also important that the observed scene has a continuous overlap across successive frames so that the camera's trajectory can be tracked consistently over time. Visual odometry is often used with other sensors, such as GPS receivers, inertial measurement units (IMUs), and LiDAR, to improve overall localization accuracy and reliability. Previous studies have demonstrated that integrating multi-directional IMUs with stereo cameras can significantly improves pose estimation precision and facilitates the robust control of unmanned aerial vehicles (UAVs) through multi-directional stereo sensing. Building upon these insights, our study adopts visual odometry based on a single IMU-stereo camera configuration. This approach emphasizes the benefits of lightweight 3D measurement systems and address the challenge of determining an accurate scale factor, which is essential for achieving reliable and practical localization performance.

In addition, in order to verify visual odometry on stairs, we minimize the brightness difference between images to ensure they align. We adopt this methodology is to improve the accuracy of self-localization estimation using visual odometry. Specifically, when a new image is observed, we solve an optimization problem that minimizes photometric error to perform the calculation. We use the inverse depth of the reference frame to restore the 3D points of each pixel and minimize the difference between the projected brightness and the brightness of the current frame. The following equation is used for optimization.

$$E_{track}(\xi) = \sum_{u \in \Omega_{D_1}} \rho\left(\frac{r_{I_{u(\xi)}}}{\sigma_{I_{r,u}}}\right) \quad (1)$$

First, the denominator represents the standard deviation of the residuals. It is calculated as the standard deviation of the predicted error obtained from the image gradient. The residuals are standardized. In other words, when the denominator value is small, the residuals are emphasized, and when the denominator value is large, the gradient is small, and the residuals are less emphasized. Finally, ρ represents the Huber function, which suppresses outliers such as brightness differences caused by exposure and lighting changes. Outliers cause the residuals to become very large and increase, squared error costs increase rapidly, which pulls the estimation results significantly and prevents the residuals from converging correctly. Specifically, we reduce the weight of pixels with large residuals so that they do not affect the estimation results adversely. Additionally, we perform 3D backprojection using depth information. Next, we will now explain the semi-global matching, which is used in depth estimation. Semi-global matching is a stereo matching algorithm that obtains disparity maps from two image pairs (left and right camera images). Advantages of semi-global matching include its robustness to noise and the ability to accurately determine distances even for objects with unclear contours.

c. Visual odometry error detection processing:

Visual odometry errors are detected by identifying discontinuities in camera pose estimation and orientation estimates, such as camera movement and camera direction on each axis. These estimates are, based on feature detection from the left and right images of the stereo camera. Basically, visual odometry errors are identified by calculating outliers in camera pose estimation and orientation estimation due to camera movement. Furthermore, we confirm that the frame numbers where the number of feature points suddenly decreases, the frame numbers where abnormal trajectories occur due to camera movement, and the frame numbers where changes in the direction of each axis occur match. Additionally, to detect frames without image blur after a sharp turn, the number of feature points is used as basis for determination.

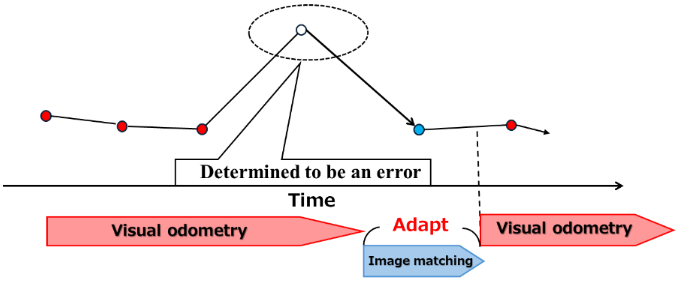

Figure 2: Visual odometry Error Recovery Methodology.

d. Visual odometry restart processing:

The restart position of visual odometry is estimated by matching feature points between the previous frame, before the error occurred, and the two subsequent frames. First, feature points are detected and extracted from consecutive images and stored in short-term memory. Next, when a visual odometry error occurs, corresponding points are obtained from the images taken before and after the error. However, since the images have low overlap rate, it is not possible to obtain the correct corresponding points. Therefore, the MSAC algorithm, an improved version of RANSAC, is used to eliminate incorrect correspondences and determine an accurate transformation matrix. When applying the MSAC algorithm, the maximum distance from a point to its corresponding point is set, and a dynamic threshold is determined based on the average matching distance. The matching success rate is calculated using Equation 1. A provisional transformation matrix is calculated to determine the correct corresponding points, and this transformation matrix is applied to all feature points. Points that match with an error below the threshold are considered correct correspondences. This procedure is repeated multiple times, and the transformation matrix with most correct corresponding points is adopted. For restart position estimation, the basic matrix is estimated using the determined corresponding points, Then the restart position is estimated using the translation vector and rotation matrix based on the relative pose from the basic matrix. Finally, camera pose estimation in visual odometry restarts using the translation vector and rotation matrix.

$$Matching\ success\ rate = \frac{Number\ of\ correct\ response\ points}{Number\ of\ corresponding\ points\ obtained} \quad (2)$$

Experiment

We conducted a walking measurement experiment on the 5th floor of the classroom building at our Toyosu campus as an indoor experimental site, as shown in Figure 4. We also performed 3D walking measurement experiments on the staircase on the 4th floor of the main building of on the same campus, as shown in Figure 5. The measurement line consisted of two straight sections: a measurement line from the starting point to a corner (14 meters long), and another from the corner to the endpoint (14 meters long). Moreover, the measurement lines on the stairs consisted of several straight lines with 11.4 meters long in total. In the experiment, we acquired temporal stereo images at 30 [fps] with a speed of 1.5 m/s using a stereo camera (ZED2, Stereolabs) with an IMU connected to a laptop PC. The stereo camera was mounted in the forward direction, parallel to the direction of motion. The stereo camera was also rotated 90 degrees horizontally at the corner to simulate a turn and measurement by an autonomous mobile robot or an indoor drone. Moreover, we evaluated error by comparing the estimated position obtained by visual odometry self-position estimation trajectory, and restart processing with the actual measured position. The position of each corner was determined by analyzing changes in the images of the scene. Visual odometry was then used to start, self-position estimation and match the length of the experimental path. For this study, the conditions for position estimation by visual odometry were based on an autonomous mobile robot and a UAV.

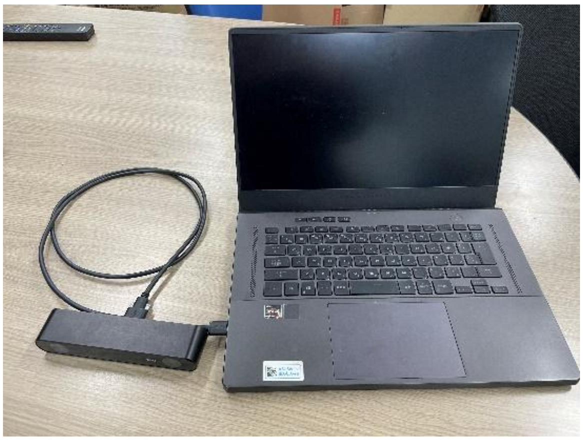

Figure 3: Experimental Site and Equipment.

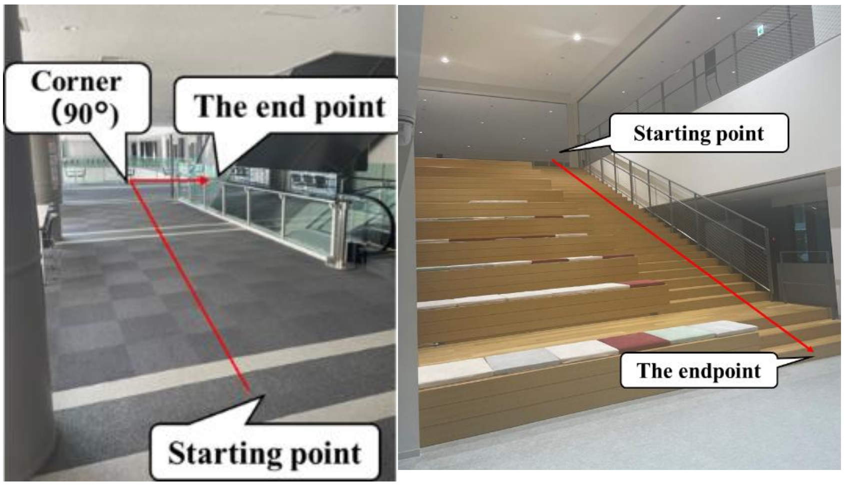

Figure 4: Experimental equipment (ZED 2 Stereo Camera and Laptop PC).

Results

a. Visual odometry restart position estimation results:

Figure 5 shows the results of the visual odometry error and restart position estimation. The following conclusions were drawn from these results. First, it was confirmed that the position at the time of the visual odometry error was (1.80, -1.32, 397.76), and that the estimated restart position was (1.91, 11.00, -0.03). When evaluating the restart position using actual measurement values, the estimation accuracy was approximately 1.91 m on the X-axis, approximately 3 m on the Y-axis, and -0.03 m on the Z-axis. In addition, regarding the possibility of recovering from camera rotation errors within a certain degree, we calculated from the rotation matrix to be 84.5 degree and 1.47 radian, Thus, we confirmed that recovery is possible by estimating the restart position within 84.5 degree and 1.47 radian.

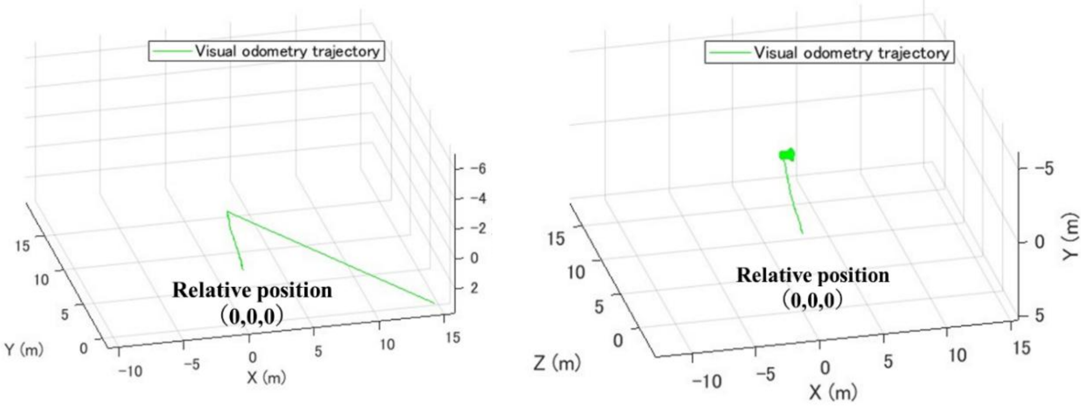

Figure 5: Visual Odometry Results (Left: Error Results, Right: Estimated Restart Position).

b. Feature point matching results:

Figures 6 and 7 show the feature point matching results obtained by the proposed methodology in this study and the feature point matching results obtained by applying RANSAC respectively. We also confirmed that the matching success rate calculated using Equation 1 increased from 3.92% to 71.43%.

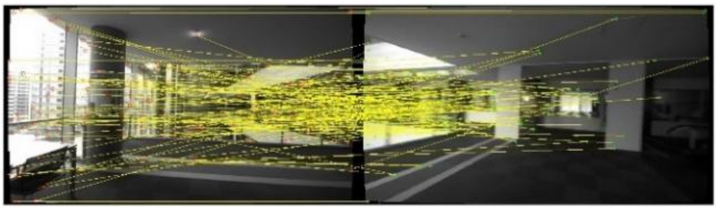

Figure 6: Feature Matching without RANSAC.

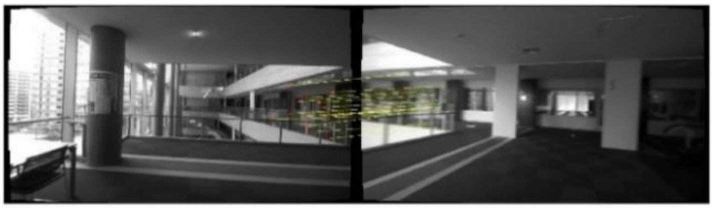

Figure 7: Feature Point Matching with RANSAC.

c. Detection of visual odometry errors:

Figure 8 shows the direction vectors of each camera posture axis calculated from the translation vector and rotation matrix by self-position estimation. Based on these results, we confirmed that the X- and Y-axes of the camera posture changed significantly, and that this posture corresponded to the position at which visual odometry errors occurred.

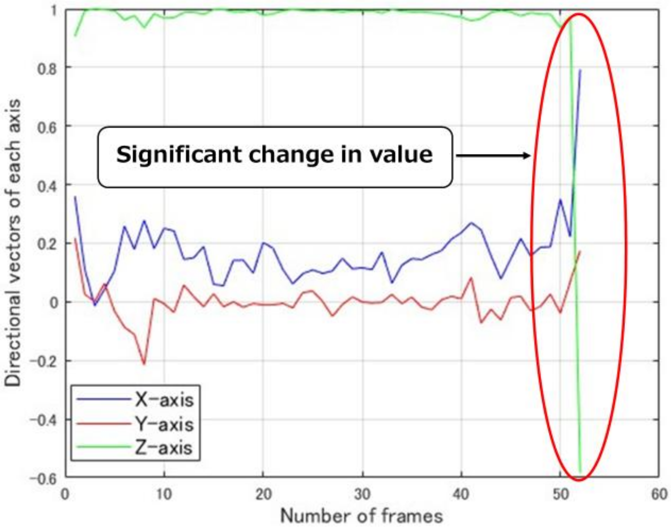

Figure 8: Directional Vector of Camera Posture for Each Axis.

Figure 9 shows the results of the number of feature points obtained from the left and right images captured by the stereo camera. I From these results, we confirmed that the number of feature points decreased sharply starting at frame number 167 in both left and right images. We also confirmed that this position at which the position where the visual odometry error occurred.

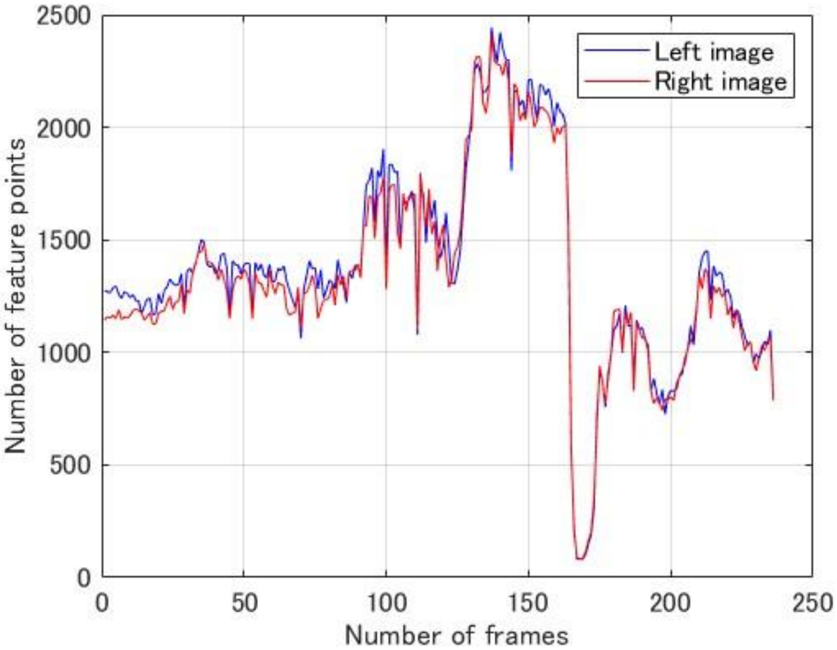

Figure 9: Number of Feature Points in Left and Right Images.

Figure 10 shows the results of camera movement for camera pose estimation using visual odometry. These results confirmed that abnormal movement amounts can be determined from visual odometry errors. We also confirmed that the restart position can be estimated using image matching and feature points.

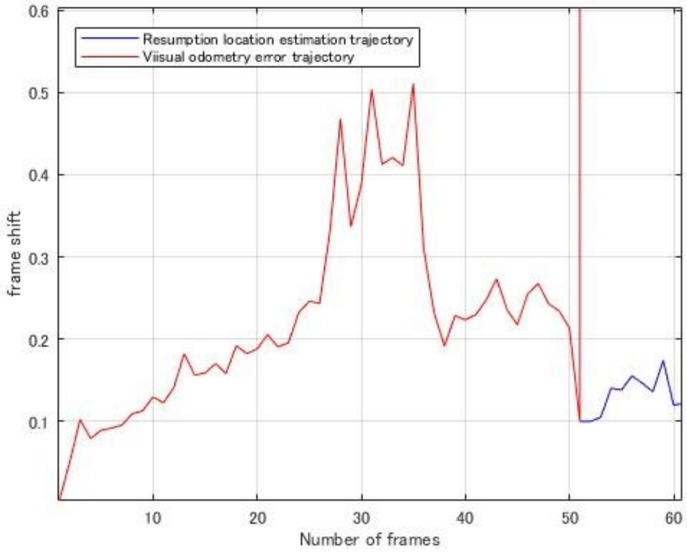

Figure 10: Camera Movement Amount.

d. Trajectory estimation results using visual odometry:

Figure 11 shows the trajectory results output from the walking measurement experiment. These results confirmed that it is possible to estimate the restart position when a visual odometry error occurs and to restart visual odometry.

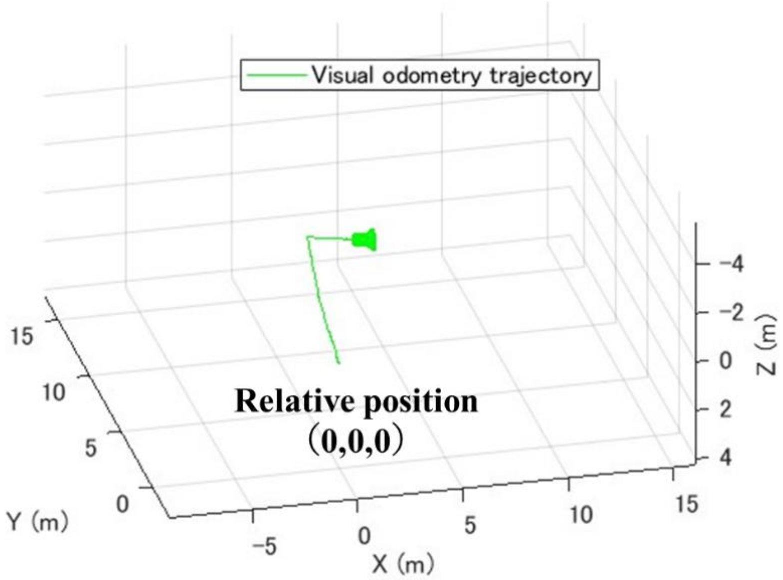

Figure 11: Trajectory Estimated by Visual Odometry.

Discussion

a. Mobile measurement experiment:

The proposed methodology in this study confirmed that visual odometer errors could be guided to the restart position (Figure 5). However, there was a significant discrepancy between the X- and Y-axis restart positions and the actual measured values. This occurred because the initial direction was not set during the experiment. It is thought that GCP installation can improve this situation. We confirmed that it is possible to calculate the corresponding points that are estimated to be correct in each image before and after the visual odometry errors using corresponding point matching based on feature points and RANSAC/MSAC, and to estimate the restart positions. In this experiment, we limited the change in direction of the movement path to right angles only, which allows us to obtain accurate corresponding points. Therefore, a future challenge is to propose a guidance processing methodology that does not depend on the direction of entry when a visual odometry error occurs.

b. Visual odometry error detection:

We confirmed that it is possible to identify the location of visual odometry errors by detecting blurring in time-series images using feature detection, as shown in Figure 12. We also confirmed that it is possible to identify frames containing visual odometry errors using feature detection. However, in some frames, the number of feature points decreased even though there were no visual odometry errors. We speculate that this was due to the dark measurement environment. Furthermore, we confirmed that visual odometry errors can be detected based on abnormal changes in camera movement values and direction vectors. Therefore, we conclude that abnormal changes in direction vectors and camera movement values can be used to detect visual odometry errors.

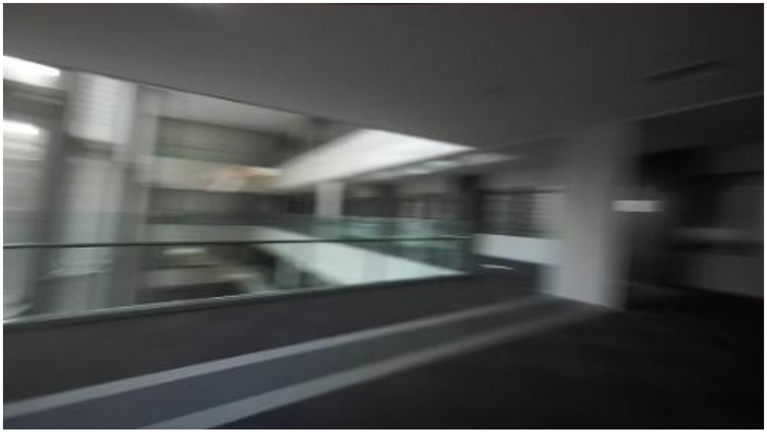

Figure 12: Example of motion blur.

c. Outlier removal using RANSAC/MSAC:

Through repeated experiments, as shown in Figure 13, applying MSAC sometimes resulted in different corresponding points. This is thought to be because due to the random feature point pairs were selected when calculating the provisional transformation matrix in MSAC. Additionally, MSAC, repeatedly evaluated random feature point pairs until an appropriate transformation matrix is determined. Therefore, variations in results may occur depending on settings such as the number of iterations or stopping conditions. Due to these two factors, different feature point pairs are selected for each trial, resulting in variability in the correspondence point matching results. Therefore, we believe that improving the randomness of MSAC with an algorithm can enhance the stability of correspondence point matching.

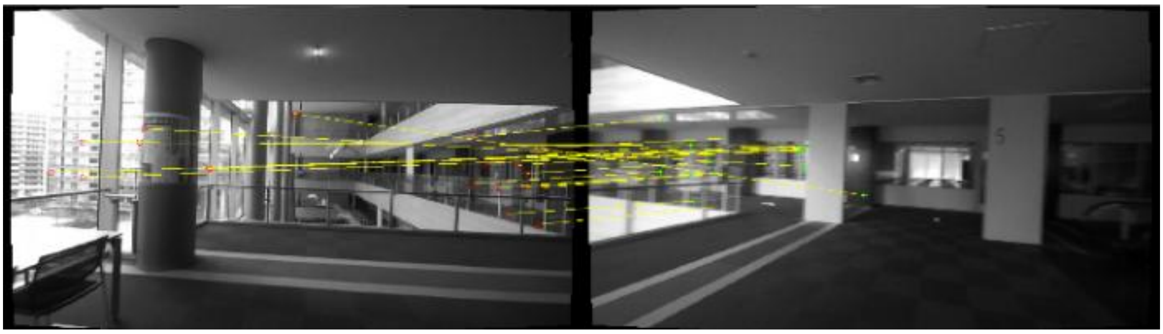

Figure13: Examples of Failures When Applying for MSAC.

d. Considerations on visual odometry on stairs:

Figure 14 shows the results of the visual odometry experiment on stairs with minimized brightness differences. These results confirm that the camera's forward position (Y-axis) position could be estimated. However, it was not possible to estimate the camera's up-down component (Z-axis) position. This is thought to be that changes in the camera's up-down position could not be distinguished from changes in its rotational position. In this experiment, the estimated results showed that the height remained almost unchanged even when the camera was moved down the stairs. This led to the interpretation that the camera was tilted forward. To address this issue, we propose combining an IMU to determine the direction of gravity, thereby enabling the coordinates to be fixed vertically. combining an IMU to determine the direction of gravity, thereby fixing the vertical coordinates.

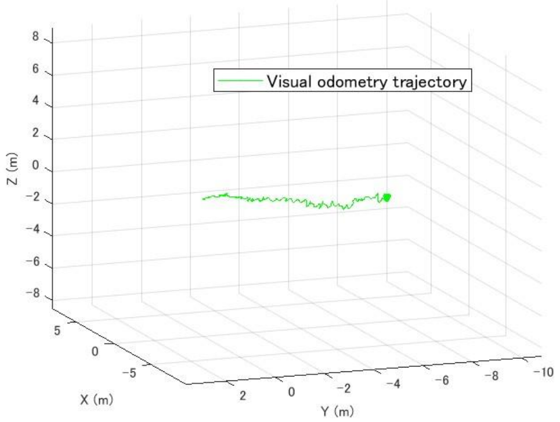

Figure 14: Visual odometry results on stairs.

Conclusion

In our study, we proposed a methodology that uses image matching on continuous images to detect visual odometry errors, guide to the restart position, and restart visual odometry. Experiments confirmed that this methodology can recover from visual odometry errors. Furthermore, we found that this the RANSAC/MSAC algorithm may produce different results when applied to this system. Future research will focus on improving this methodology to estimate a transformation matrix independent of the MSAC algorithm, as the randomness of the current MSAC algorithm results in unstable correspondence point matching. Our improvement plan focuses on constructing an algorithm that mitigates the randomness of the MSAC algorithm. We also plan to address technical issues such as robustness to lighting conditions and motion blur at corners. In addition, we plan to conduct verification experiments using a camera combined with an IMU to minimize brightness differences and perform visual odometry in locations where identifying landmarks is difficult, such as on the lunar surface. Future challenges include verifying the proposed methodology's functionality through 3D motion experiments and developing transformation matrix estimation methodology independent of the MSAC algorithm.

Reference

Zhang Z., (2000). A flexible new technique for camera calibration, IEEE, pp.1330-1334.

Lucas R, Nuno M. R., Maria I. P., Antoine H., Andry M. P., (2022). A Practical Survey on Visual Odometry for Autonomous Driving in Challenging Scenarios and Conditions, IEEE, pp.72182-72205.

Handa A., Ehelan T., McDonald J., (2014). Benchmark for RGB-D Visual odometry, 3Dreconstruction and SLAM, IEEE, pp.1050-4729.

IImir Z. I., IIya M. A., (2017). Comparison of ROS-based visual SLAM methods in homogeneous indoor environment, 14th Workshop on Positioning, Navigation and Communications(WPNC).

Zachary T., Lahav L., Jia D., (2023). DEEP Patch Visual Odometry, NeurlPS.

Zachary T., Jia D., (2021). DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras, Zachary Teed, Jia Deng, NeurlPS.

Chaolei, W., Tianmiao, W., Jianhong, L., Yang, C., Yicheng, Z., Cong, (2012). Monocular visual SLAM for small UAVs in GPS-denied environments, IEEE.

Saito K., Ozaki G., Okudaira K., Nakagawa M., (2022). Performance Verification Of Visual Odometry with IMU-Stereo Camera in Indoor UAV, The 16th South East Asian Technical University Consortium Symposium, pp.27-30.

Alexander K., Markus B., Yogesh K., (2020). Real-time Mapping for Unmanned Aerial Vehicles, Ulf Bestmann International Conference on Unmanned Aircraft Systems(ICUAS).

Ke, S, Kartik, M, Bernd, P, Michael, W, Sikang, L, and Yash, M, (2018). Robust Stereo Visual Intertial Odometry for Fast Autonomous Flight, IEEE, pp.965-972.

Faragher, R, M., Harle, R, K, Nashville, (2013). SmartSLAM - An Efficient Smartphone Indoor Positioning System Exploiting Machine Learning and Opportunistic Sensing, pp.1006-1019.

Saito K., Nakagawa M., Kawasaki Y., Takebayashi M., Miwa M., (2024). Verification of Indoor-outdoor Seamless Positioning and Camera Positioning Rectification Systems Mounted on Infrastructure Inspection UAVs, The 18th South East Asian Technical University Consortium Symposium, pp.81-84.

Zachary T., Jia D., (2020). Video To Depth With Differentiable Structure From Motion, ICLR, DEEPV2D, Computer Science Center for Statistics & Machine Learning Princeton Language and Intelligence (PLI).

Jakob E., Jorg S., Daniel C. (2015). Large-scale direct SLAM with stereo cameras, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

Jakob E., Thomas S., Daniel C. (2014), LSD-SLAM: Large-scale direct monocular SLAM, European Conference on Computer Vision, LNIP, volume 8690

Eshima Y., Saito K., Nakagawa M., (2024). Trajectory Recovery of Visual Odometry at Coners by Temporal Stereo Point Cloud Registration, Asia Conference on Remote Sensing, 17 pages.